Statistical Papers (2024) 65:381–413
https://doi.org/10.1007/s00362-023-01398-0

REGULAR ARTICLE

# Generalised score distribution: underdispersed continuation of the beta-binomial distribution

Bogdan Ćmiel[1] · Jakub Nawała[2] · Lucjan Janowski[2] · Krzysztof Rusek[2]

Received: 12 July 2022 / Accepted: 24 January 2023 / Published online: 9 February 2023

## Abstract

Consider a class of discrete probability distributions with a limited support. A typical example of such support is some variant of a Likert scale, with a response mapped to either the $\{1, 2, \ldots, 5\}$ or $\{-3, -2, \ldots, 2, 3\}$ set. Such type of data is common for Multimedia Quality Assessment but can also be found in many other research fields. For modelling such data a latent variable approach is usually used (e.g., Ordered Probit). In many cases it is convenient or even necessary to avoid latent variable approach (e.g., when dealing with too small sample size). To avoid it the proper class of discrete distributions is needed. The main idea of this paper is to propose a family of discrete probability distributions with only two parameters that play the same role as the parameters of the normal distribution. We call the new class the Generalised Score Distribution (GSD). The proposed GSD class covers the entire set of possible means and variances, for any fixed and finite support. Furthermore, the GSD class can be treated as an underdispersed continuation of a reparametrized beta-binomial distribution. The GSD class parameters are intuitive and can be easily estimated by the method of moments. We also offer a Maximum Likelihood Estimation (MLE) algorithm for the GSD class and evidence that the class properly describes response distributions coming from 24 Multimedia Quality Assessment experiments. At last, we show that the GSD class can be represented as a sum of dichotomous zero–one random variables, which points to an interesting interpretation of the class.

**Keywords** Extension of beta-binomial distribution · Overdispersion · Underdispersion · Multimedia quality assessment · Discrete probability distribution · Likert Scale

**Mathematics Subject Classification** 62P30 · 62F10 · 62F03 · 62F40

✉ Bogdan Ćmiel
cmielbog@gmail.com

[1] Department of Applied Mathematics, AGH University of Science and Technology, Krakow, Poland

[2] Institute of Telecommunications, AGH University of Science and Technology, Krakow, Poland

## 1 Introduction

A Likert scale is used in numerous research fields, like psychology, medicine, or quality of experience (Liddell and Kruschke 2018; Pinson and Janowski 2014), to name a few. Responses given to questions using a Likert scale are ordinal, but in practice they are often analysed as data coming from an interval scale. There is even a fairly common approach of ignoring the fact that data are discrete and treating them as if they come from a continuous, usually normal, distribution (ITU-R 2019). The reason why ordinal responses are converted to an interval scale is that proper latent variable models, like ordered logit or ordered probit [see McCullagh and Nelder (1989)], are too complicated for studies with relatively few responses per hidden variable. The parameters in hidden variable models are difficult to interpret and in the case of small sample sizes, it is impossible to estimate them properly.

The main idea of this paper is to propose a family of discrete probability distributions that is able to model responses gathered in Multimedia Quality Assessment (MQA)[1] subjective experiments, and we believe it can be used in other fields. In MQA the distribution of responses is often (but not exclusively) uderdispersed (i.e., having variance lower than that of the binomial distribution) and practitioners analysing these data need a model intuitively conveying response distribution characteristics. We put forward a solution avoiding the difficulties inherent to treating ordinal responses as though they are expressed on an interval scale and, at the same time, offer a model with only two parameters (to sidestep overparameterisation challenges). We find our solution attractive to practitioners, who want to use relatively simple but mathematically correct tools. We call our proposed family of distributions a *Generalised Score Distribution* class (GSD). It is a two-parameter family of discrete distributions on the set $\{1, \dots, M\}$, $M \in \mathbb{N} \backslash \{1, 2\}$. Because of its convenient parameterisation, the class does a similar job to what the normal distribution class does for the continuous case. The first parameter $\psi \in [1, M]$ is (as in the case of the normal distribution class) the expected value. We would like the second parameter of the GSD class to play the same role as the second parameter of the normal distribution class. Unfortunately, for discrete distributions defined on the set $\{1, \dots, M\}$, the range of all possible variances is changing with $\psi$. Therefore, the second parameter $\rho \in [0, 1]$ (also referred to as *dispersion parameter* or *confidence parameter*) of the GSD class is a linear function of variance, with the value range equal to the interval [0, 1] for every $\psi$. To the best of our knowledge, the convenient parameterisation of the GSD class is unique, compared to other discrete distributions. It allows us to look at the parameters describing the expected value and variance independently, as one can do in the case of the normal distribution class. In other words, shifting the expected value parameter ($\psi$), does not change the dispersion parameter ($\rho$). Likewise, changing the dispersion parameter ($\rho$), does not influence the expected value parameter ($\psi$). The GSD class covers all possible first- and second-order moments for discrete distributions defined on $\{1, \dots, M\}$. It also offers explicit formulae for probabilities without the use of special functions and thus explicit formulae for the derivatives of its log-likelihood function. Therefore, we believe that the GSD can be used outside of the field of MQA.

[1] Video Quality Assessment (VQA) mentioned previously is a sub-field of MQA.

The problem of generalising the binomial distribution to overdispersed and underdispersed data is known and widely considered in literature. The natural generalisation of the binomial distribution for overdispersed data is the beta-binomial distribution. It provides simple formulae for probabilities and derivatives of its log-likelihood function. Furthermore, there are available ready-to-use algorithms for the estimation of its parameters (see Griffiths (1973)). One can also find applications of the beta-binomial distribution when dealing with overdispersed data (see e.g. Gange et al. (1996)). In Prentice (1986) the method for extending the beta-binomial distribution for underdispersed data is proposed. Unfortunately, this method does not provide a distribution that covers all possible variances. For any fixed mean value, it stops at some variance and cannot go lower (see Fig. 2). Thus, the data that are strongly underdispersed cannot be modelled in such a way. In this paper, we propose a different way of extending the beta-binomial distribution. Our solution allows to obtain all possible variances for a discrete distribution defined on $\{1, \ldots, M\}$. Our method also provides a reparameterisation to obtain easy to interpret parameters $\psi$ (mean value) and $\rho$ (confidence level linearly dependent on variance). The GSD class can be also represented as a sum of dichotomous zero–one random variables (see Proposition 4), which gives an interesting interpretation of that class.

This paper provides estimation and test of goodness-of-fit algorithms for the GSD class. We analyse the algorithms performance through an extensive simulation study. We also use six Multimedia Quality Assessment (MQA) databases (amounting to more than 100,000 individual responses) to provide evidence that the proposed class is a useful analytical tool that can be used in practice. It is worth mentioning that we already proposed in the past a tool based on the GSD class. The tool extends possible ways to validate data consistency of responses obtained during a MQA subjective experiment (Nawała et al. 2020). Importantly, our work was noticed by practitioners in the MQA field and referred to in Chinen et al. (2021), Chinen (2021), and Hoßfeld et al. (2021).

The GSD class we propose can be used for modelling survey responses expressed on a Likert scale as well. For example, in Alwin et al. (2018) the authors consider the impact of the number of response categories on the reliability of the measurements. In the theorised response generation model (provided in equation (1) of Alwin et al. (2018)), one can use the GSD class to model the random error. This approach would then allow to estimate the latent unobserved true response. The same method can also be used, for example, in Malott et al. (2017), where the problem of measuring patient experience in hospitals is considered.

We claim that the GSD class properly describes responses from MQA subjective experiments. We also state that the GSD class estimates well response distributions even for samples of small size (i.e., sample sizes conventionally used in MQA experiments). At last, we argue that the GSD class describes response distributions using easy to interpret and easy to estimate parameters. To substantiate our claims, in this work we present the following contributions:

- We offer the GSD family of distributions (also referred to as the GSD class) that:
  1. covers all possible first- and second-order moments for a distribution defined on a discrete finite support,

2. extends binomial distribution to cover all underdispersed and overdispersed data, and
3. uses parameterisation similar to normal distribution's parameterisation.

- We evidence that the GSD family of distributions can be represented as a sum of dichotomous zero–one random variables.
- We show a Maximum Likelihood Estimation (MLE) algorithm for the GSD class.
- We indicate through goodness-of-fit testing that the GSD class well describes responses from MQA subjective experiments (in contrast to other commonly used modelling approaches).
- We reveal that based on samples of small size, the GSD class better forecasts a response distribution for a sample of larger size, in comparison to the empirical distribution.

The paper is structured as follows. In Sect. 2, we describe the proposed GSD family of distributions and compare it with the Ordered Probit model. Section 3 introduces the maximum likelihood estimation for the GSD class. It also considers the numerical accuracy of the estimation method we use. (Numerical accuracy was also examined in multidimensional case in Appendix E.) Sect. 4 presents the analyses performed on real data sets of responses from six MQA studies. The last section concludes the paper. All proves and additional formulae can be found in Appendices.

## 2 Model description

Assuming that we use an $M$-point discrete scale, a random variable $U$ (describing a subjective response) has a distribution given by:

$$P(U = s) = p_s, \text{ where } \sum_{s=1}^{M} p_s = 1 \tag{1}$$

Such a description of a response distribution is general but has $M-1$ different parameters. There are $M-1$ of them, since there are $M$ probabilities this distribution describes. In general, we can describe a subjective response as a function:

$$U = \psi + \epsilon, \tag{2}$$

where $\psi$ is the expected value (referred to as true quality[2] in the context of MQA research) and $\epsilon$ is an error term with the mean value equal to zero. An algorithm predicting stimulus quality (or any other subjectively judged trait) should aim at estimating $\psi$. Still, the error distribution is important and should be modelled. The error term represents the precision of $\psi$ estimation. It is desirable that the error term (represented by $\epsilon$) should not be too complicated. Therefore, we would like to use a model in which the error is described by a single parameter. (Please note that in Appendix E

[2] Our notation convention generally follows the guidelines of VQEG (Video Quality Expert Group), described in Janowski et al. (2019).

we also consider a multidimensional version of subjective responses with $m$ quality parameters and $n$ error parameters.)

### 2.1 Ordered probit with fixed thresholds

The models proposed by Janowski and Pinson (2015) and Li and Bampis (2017) describe subjective responses as following a continuous normal distribution with certain mean $\mu$, which is assumed to represent the latent stimulus quality (also referred to as true quality), and standard deviation $\sigma$, describing the error. Therefore, subjective response $O \sim \mathcal{N}(\mu, \sigma^2)$. Since in MQA experiments subjective responses are often expressed on a discrete scale, we cannot directly observe $O$. To convert the continuous form of a response to a discrete one, discretisation and censoring (clipping) are necessary. This process converts a continuous random variable $O$ to a discrete variable $U$. We can calculate each response category probability (i.e., $U$ distribution), as a function of $\mu$ and $\sigma$ using the following equations:

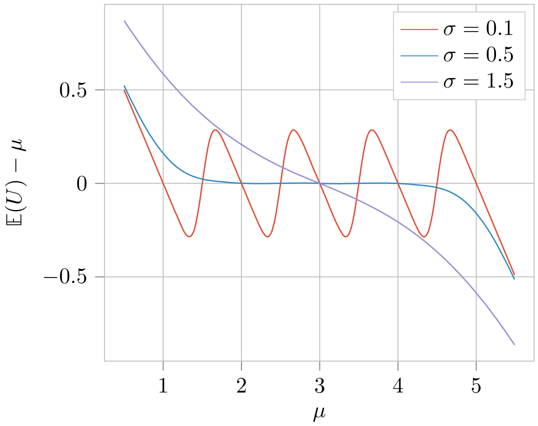
$$P(U = s) = \int_{s-0.5}^{s+0.5} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \tag{3}$$

for $s = \{2, 3, \ldots, M-1\}$ and

$$\begin{aligned} P(U = 1) &= \int_{-\infty}^{1.5} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx, \\ P(U = M) &= \int_{M-0.5}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx. \end{aligned} \tag{4}$$

Note that the definition of true quality $\mu$ is model dependent here. This is a flaw of this approach. It is also worth mentioning that $\mu = \mathbb{E}(O)$ can be completely different from $\psi = \mathbb{E}(U)$. The latter is a natural and model-independent parameter defining the true quality (for possible differences between $\mathbb{E}(O)$ and $\mathbb{E}(U)$ see Fig. 1).

**Fig. 1** The difference between $\mathbb{E}(U)$ and $\mu$ (the ordered probit parameter) for $M = 5$ and different $\sigma$



It is obvious that $\mathbb{E}(O) \in (-\infty, \infty)$ and $\mathbb{E}(U) \in [1, M]$ have to be different. However, even for $\mathbb{E}(O) \in [1, M]$ the differences are considerable. The other problem, especially from the numerical estimation point of view, is the unbounded parameter set $(\mu, \sigma) \in (-\infty, \infty) \times (0, \infty)$. In Fig. 6 one can see how parameters $(\mu, \sigma)$ map to $(\mathbb{E}(U), \mathbb{V}(U))$ after discretisation, for the case $M = 5$. This figure and the lack of an inverse formula for calculating $(\mu, \sigma)$ having $(\mathbb{E}(U), \mathbb{V}(U))$, make it clear that this approach (referred to in the literature as *ordered probit* (Becker and Kennedy 1992)) may prove problematic if the moments-based estimation would be used. Likewise, it may be challenging to find a starting point for numerical estimation methods.

## 2.2 GSD

An example of a discrete distribution that is described by two parameters is the beta-binomial distribution (Coombes 2018). The smallest variance the beta-binomial distribution can express is binomial distribution's variance. It is a strong limitation. For example, in the case of MQA subjective experiments for video, in Hossfeld et al. (2018) it is suggested that the binomial distribution has the highest possible variance for a correctly conducted subjective experiment. Differently put, most MQA subjective experiments yield data with response distributions having the variance lower than that of the binomial distribution. Therefore, we need a different distribution, covering the whole spectrum of possible variances. This is especially true for underdispersed probability distributions (i.e., distributions with the variance lower than that of the binomial distribution; see Fig. 2).

### 2.2.1 GSD construction and definition

Let us start with the equation describing subjective responses

$$U = \psi + \epsilon, \tag{5}$$

where $\epsilon$ is an error with mean value equal to 0. Since $U$ belongs to the set $\{1, 2, \ldots, M\}$, then the distribution of $\epsilon$ has to be supported on the set $1 - \psi$, $2 - \psi$,..., $M - \psi$. Let us consider the shifted binomial distribution for $\epsilon$:

$$P(\epsilon = k - \psi) = \binom{M-1}{k-1} \left(\frac{\psi - 1}{M - 1}\right)^{k-1} \left(\frac{M - \psi}{M - 1}\right)^{M-k},$$

where $k \in \{1, \ldots, M\}$ represents the response categories, from which subjective experiment participants (also referred to as *subjects* or *raters*) can choose from.

Since the support of this distribution and the mean value are fixed, we obtain a fixed shifted binomial distribution without any freedom. However, we would like to have a class of distributions for $\epsilon$ with all possible variances $\mathbb{V}(\epsilon) = \mathbb{V}(U)$. Let us think about how the set of all possible variances, for all distributions supported on the set $1 - \psi$, $2 - \psi$,..., $M - \psi$, looks like. Remember that the mean values for such distributions are fixed at 0 (since they describe the error term). This is why the set of

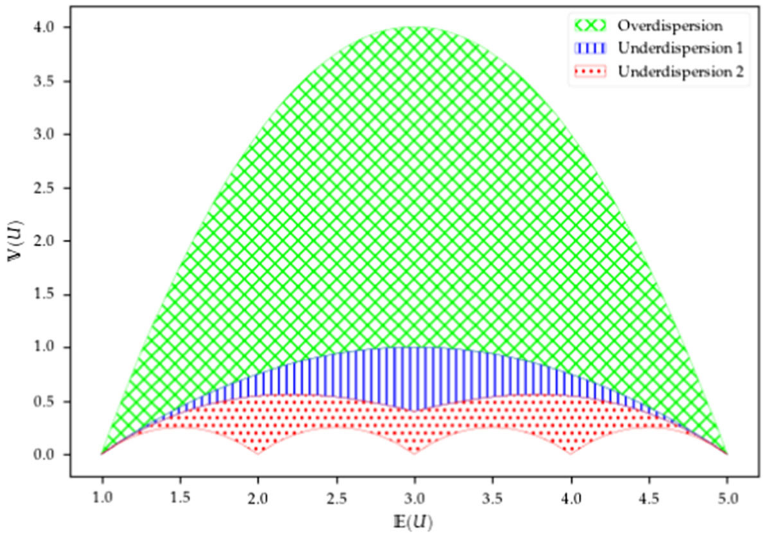


**Fig. 2** Area of all possible $(\mathbb{E}(U), \mathbb{V}(U))$ for discrete distributions on $\{1, \ldots, M\}$ for $M = 5$. The area is divided into different parts covered by two-parameter discrete models: the beta-binomial distribution (green crosses), extended beta-binomial distribution (previous + blue vertical hatching), and our solution (covers all possible values including the area marked with red dots)

all possible variances depends on $\psi$. If we denote by $V_{\min}(\psi)$, $V_{\max}(\psi)$ the minimal and maximal possible variance, respectively, then

$$V_{\min}(\psi) = (\lceil \psi \rceil - \psi)(\psi - \lfloor \psi \rfloor), \tag{6}$$

$$V_{\max}(\psi) = (\psi - 1)(M - \psi), \tag{7}$$

and the interval $[V_{\min}(\psi), V_{\max}(\psi)]$ is the set of all possible variances. Notice that the interval $[V_{\min}(\psi), V_{\max}(\psi)]$ is the biggest for $\psi = (M + 1)/2$ and if $\psi$ is not an integer, then $V_{\min}(\psi) > 0$. Let us return to the shifted binomial distribution. It is easy to calculate that its variance is equal to:

$$V_{\mathrm{Bin}}(\psi) := \frac{V_{\max}(\psi)}{M - 1}. \tag{8}$$

The question is how to obtain from this shifted binomial distribution a class of distributions that covers the whole interval of variances $[V_{\min}(\psi), V_{\max}(\psi)]$ for any $\psi$ (see Fig. 2).

We would like to obtain this class by:

- adding only a single normalised parameter $\rho \in [0, 1]$,
- making variance of the error to be linearly dependent on $\rho$ and
- requiring that variance is a decreasing function of $\rho$ (this way we could interpret $\rho$ as a confidence parameter; see Fig. 4).

Let us denote by $H_\rho$ the distribution of the error fulfilling the above conditions. Since the variance of the error is linearly dependent on $\rho$ and decreasing, then it has to be equal to

$$V_{H_\rho}(\epsilon) = \rho V_{\min}(\psi) + (1-\rho) V_{\max}(\psi). \tag{9}$$

Using formulae (8) and (9) we can calculate

$$V_{H_\rho}(\epsilon) = V_{\mathrm{Bin}}(\psi) \;\Leftrightarrow\; \rho = C(\psi) := \frac{M-2}{M-1} \frac{V_{\max}(\psi)}{V_{\max}(\psi) - V_{\min}(\psi)}, \tag{10}$$

which gives us the value of $\rho$ corresponding to a shifted binomial distribution. We have

$$V_{H_\rho}(\epsilon) \in [V_{\min}(\psi), V_{\mathrm{Bin}}(\psi)] \Leftrightarrow \rho \in [C(\psi), 1],$$

which corresponds to the red coloured dots and blue vertical hatching in Fig. 2, and

$$V_{H_\rho}(\epsilon) \in [V_{\mathrm{Bin}}(\psi), V_{\max}(\psi)] \Leftrightarrow \rho \in [0, C(\psi)],$$

which corresponds to the green coloured area in Fig. 2.

For variances bigger than $V_{\mathrm{Bin}}(\psi)$ (the green coloured area in Fig. 2) we use the reparameterised beta binomial distribution. Since the mean value is fixed, we only have one free parameter $\rho \in [0, C(\psi)]$. The effect of such reparameterization gives us the distribution denoted by $G_\rho$:

$$\begin{aligned} & P_{G_\rho}(\epsilon = k - \psi) \\ & = \binom{M-1}{k-1} \frac{\prod_{i=0}^{k-2} \left( \frac{(\psi-1)\rho}{(M-1)} + i(C(\psi)-\rho) \right) \prod_{j=0}^{M-k-1} \left( \frac{(M-\psi)\rho}{(M-1)} + j(C(\psi)-\rho) \right)}{\prod_{i=0}^{M-2} (\rho + i(C(\psi)-\rho))}, \end{aligned} \tag{11}$$

where $\rho \in [0, C(\psi)]$ and $k \in \{1, \ldots, M\}$. The above formula can be rewritten as

$$\begin{aligned} & P_{G_\rho}(\epsilon = k - \psi) \\ & = \begin{cases} \frac{M-\psi}{M-1} \prod_{i=1}^{M-2} \frac{\frac{(M-\psi)\rho}{(M-1)} + i(C(\psi)-\rho)}{\rho + i(C(\psi)-\rho)} \text{ for } k = 1 \\ \binom{M-1}{k-1} \frac{(\psi-1)(M-\psi)\rho}{(M-1)^2} \frac{\prod_{i=1}^{k-2} \left( \frac{(\psi-1)\rho}{(M-1)} + i(C(\psi)-\rho) \right) \prod_{j=1}^{M-k-1} \left( \frac{(M-\psi)\rho}{(M-1)} + j(C(\psi)-\rho) \right)}{\prod_{i=1}^{M-2} (\rho + i(C(\psi)-\rho))} \text{ for } \\ \qquad k = 2, \ldots, M-1 \\ \frac{\psi-1}{M-1} \prod_{i=1}^{M-2} \frac{\frac{(\psi-1)\rho}{(M-1)} + i(C(\psi)-\rho)}{\rho + i(C(\psi)-\rho)} \text{ for } k = M \end{cases} \end{aligned} \tag{12}$$

Therefore, for $\rho = 0$ we obtain

$$P_{G_0}(\epsilon = k - \psi) = \begin{cases} \frac{M-\psi}{M-1} & \text{for } k = 1 \\ 0 & \text{for } k = 2, \dots, M-1 \\ \frac{\psi-1}{M-1} & \text{for } k = M \end{cases}$$

**Proposition 1** *If $\epsilon$ has $G_\rho$ distribution for fixed $\psi \in [1, M]$ and $\rho \in [0, C(\psi)]$, then*

$$\mathbb{E}(U) = \psi, \;\; \mathbb{V}(U) = \rho V_{\min}(\psi) + (1-\rho)V_{\max}(\psi), \;\; \mathbb{V}(U) \in [V_{\mathrm{Bin}}(\psi), V_{\max}(\psi)],$$

*where $U = \psi + \epsilon$ is supported on $\{1, \dots, M\}$.*

The proof of Proposition 1 can be found in Appendix A.

***Remark 1*** Notice that for $\rho \to 0$ the $G_\rho$ distribution approaches a two-point distribution supported on $\{1-\psi, M-\psi\}$, with the biggest possible variance equal to $V_{\max}(\psi)$. For $\rho \to C(\psi)$ the $G_\rho$ distribution approaches the shifted binomial distribution, with variance equal to $V_{\mathrm{Bin}}(\psi)$.

For variances smaller than $V_{\mathrm{Bin}}(\psi)$ (cf. the blue vertical hatching and red coloured dots in Fig. 2) we use a mixture technique. Specifically, we take a mixture of the shifted binomial distribution and the distribution with the smallest possible variance (i.e., a two-point or one-point distribution, depending on $\psi$). Of course, the mixture parameter has to be reparameterised to fit the $[C(\psi), 1]$ interval. The effect of such reparameterisation gives us the distribution denoted by $F_\rho$:

$$P_{F_\rho}(\epsilon = k - \psi) \\ = \frac{\rho - C(\psi)}{1 - C(\psi)}[1 - |k - \psi|]_+ + \frac{1-\rho}{1-C(\psi)}\binom{M-1}{k-1}\left(\frac{\psi-1}{M-1}\right)^{k-1}\left(\frac{M-\psi}{M-1}\right)^{M-k}, \tag{13}$$

where $\rho \in [C(\psi), 1]$, $[x]_+ = \max(x, 0)$ and $k \in \{1, \dots, M\}$.

**Proposition 2** *If $\epsilon$ has $F_\rho$ distribution for fixed $\psi \in [1, M]$ and $\rho \in [C(\psi), 1]$, then*

$$\mathbb{E}(U) = \psi, \;\; \mathbb{V}(U) = \rho V_{\min}(\psi) + (1-\rho)V_{\max}(\psi), \;\; \mathbb{V}(U) \in [V_{\min}(\psi), V_{\mathrm{Bin}}(\psi)],$$

*where $U = \psi + \epsilon$ is supported on $\{1, \dots, M\}$.*

The proof of Proposition 2 can be found in Appendix A.

***Remark 2*** Notice that for $\rho \to C(\psi)$ the distribution $F_\rho$ approaches the shifted binomial distribution with variance equal to $V_{\mathrm{Bin}}(\psi)$. For $\rho \to 1$ the distribution $F_\rho$ approaches a two-point or one-point distribution (depending on $\psi$), with the smallest possible variance equal to $V_{\min}(\psi)$.

Finally, we obtain the distribution:

$$H_\rho = G_\rho \, I(\rho < C(\psi)) + F_\rho \, I(\rho \geq C(\psi)).$$

**Definition 1** If $\epsilon$ has $H_\rho$ distribution for fixed $\psi \in [1, M]$ and $\rho \in [0, 1]$, then we say that $U = \psi + \epsilon$ has the GSD($\psi, \rho$) distribution, where $\psi$ is the expected value and $\rho \in [0, 1]$ is a confidence parameter, linearly dependent on the variance, i.e.,

$$\rho = \frac{V_{\max}(\psi) - \mathbb{V}(U)}{V_{\max}(\psi) - V_{\min}(\psi)}$$

**Proposition 3** *If $U$ supported on $\{1, \ldots, M\}$ has the GSD($\psi, \rho$) distribution for $\psi \in [1, M]$ and $\rho \in [0, 1]$, then*

$$\mathbb{E}(U) = \psi, \ \ \mathbb{V}(U) = \rho V_{\min}(\psi) + (1-\rho)V_{\max}(\psi), \ \ \mathbb{V}(U) \in [V_{\min}(\psi), V_{\max}(\psi)].$$

The proof of Proposition 3 is an obvious consequence of Propositions 1 and 2.

In the following proposition, we show that our GSD class can be looked at from an interesting angle. The GSD class can be represented as a distribution of the number of successes in a sequence of $M - 1$ experiments.

**Proposition 4** *GSD distribution can be represented as a sum of dichotomous zero–one random variables. Specifically, $U = 1 + \sum_{i=1}^{M-1} Z_i$ has the GSD($\psi, \rho$) distribution if:*

(a) *in the case of $\rho \geq C(\psi)$, $Z_1, \ldots, Z_{M-1}$ are zero–one independent random variables and*

$$P(Z_i = 1) = \phi_{\psi,\rho}(i),$$

*where* (*see Fig.* 3)

$$\phi_{\psi,\rho}(x) = \begin{cases} \frac{\rho - C(\psi)}{1-C(\psi)} + \frac{(1-\rho)(\psi-1)}{(1-C(\psi))(M-1)} & \text{for } x \leq \psi - 1 \\ \frac{\rho - C(\psi)}{1-C(\psi)}(\psi - x) + \frac{(1-\rho)(\psi-1)}{(1-C(\psi))(M-1)} & \text{for } x \in (\psi - 1, \psi] \\ \frac{(1-\rho)(\psi-1)}{(1-C(\psi))(M-1)} & \text{for } x > \psi \end{cases} .$$

(b) *in the case of $\rho < C(\psi)$, $Z_1|B, \ldots, Z_{M-1}|B$ are conditionally independent zero–one random variables, where $P(Z_i = 1|B) = B$ and $B$ has the beta distribution of the following form: $\mathcal{B}\left(\frac{(\psi-1)\rho}{(M-1)(C(\psi)-\rho)}, \frac{(M-\psi)\rho}{(M-1)(C(\psi)-\rho)}\right)$.*

The proof of Proposition 4 can be found in Appendix A.

***Remark 3*** If we consider a class of functions $\phi_{\psi,\rho}$ satisfying the following conditions:

- $\forall\, x \in [1, M-1] \ \ \phi_{\psi,\rho}(x) \in [0, 1]$,
- $\sum_{i=1}^{M-1} \phi_{\psi,\rho}(i) = \psi - 1$,
- $\sum_{i=1}^{M-1} [\phi_{\psi,\rho}(i)]^2 = \psi - 1 - (1-\rho)V_{\max}(\psi) - \rho V_{\min}(\psi)$,

then we obtain a completely general response distribution suitable for representing underdispersed data (cf. Fig. 2) with mean value $\psi$ and confidence parameter $\rho$.

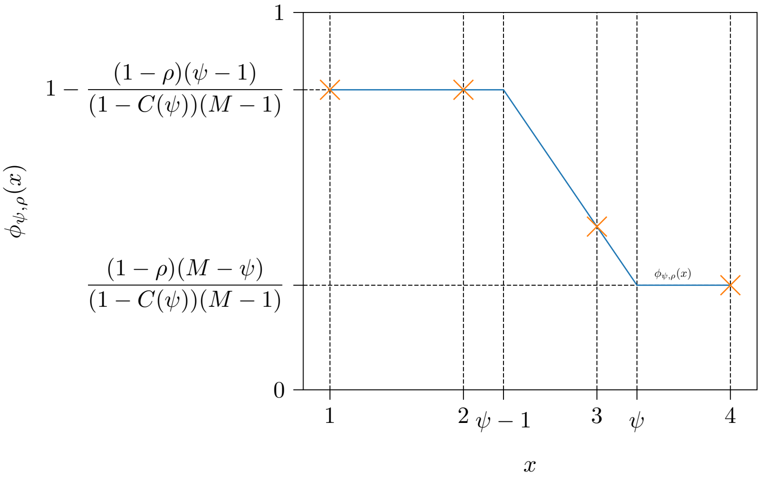
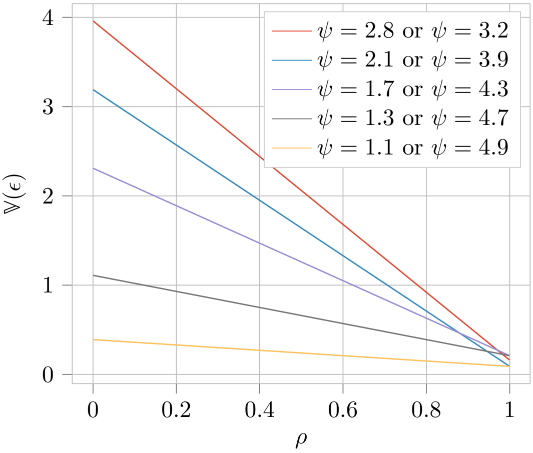
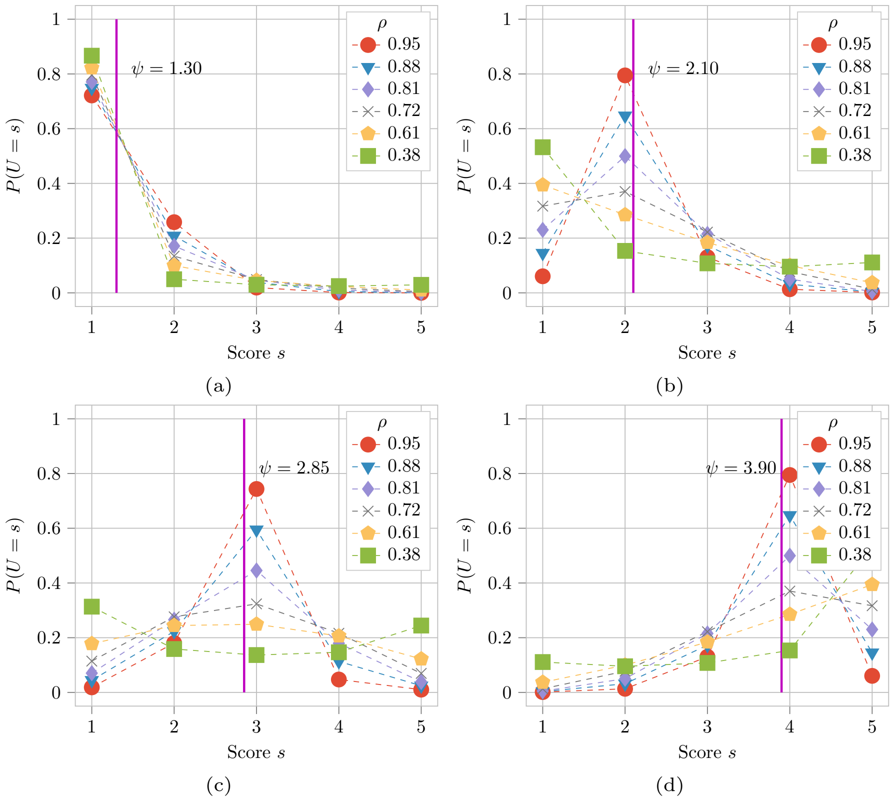


**Fig. 3** Example of $\phi_{\psi,\rho}(x)$ for $\psi = 3.3$, $\rho = 0.9$, and $M = 5$

**Fig. 4** Variance of the error term $\epsilon$ for $M = 5$



The motivation behind constructing the GSD class is to have a class that would properly describe response distributions observed when analysing responses from MQA subjective experiments. There, the responses are often expressed on a 5-level scale, with the following mapping between discrete consecutive numbers and textual labels: 1—Bad, 2—Poor, 3—Fair, 4—Good and 5—Excellent. Although initially designed for the MQA research, we expect the GSD class to be useful for describing other, more general processes (at least for measurement processes exhibiting a characteristic similar to what is the case for the MQA subjective experiments) (Fig. 4). Examples of specific incarnations of the GSD family of distributions (for $M = 5$) are shown in Fig. 5. Note that for $\psi$ close to 1 or $M$, regardless of $\rho$, the obtained distributions are similar. This is because the maximum spread we can obtain is limited by a small range of possible variances (see Fig. 2).



(a)

(b)

(c)

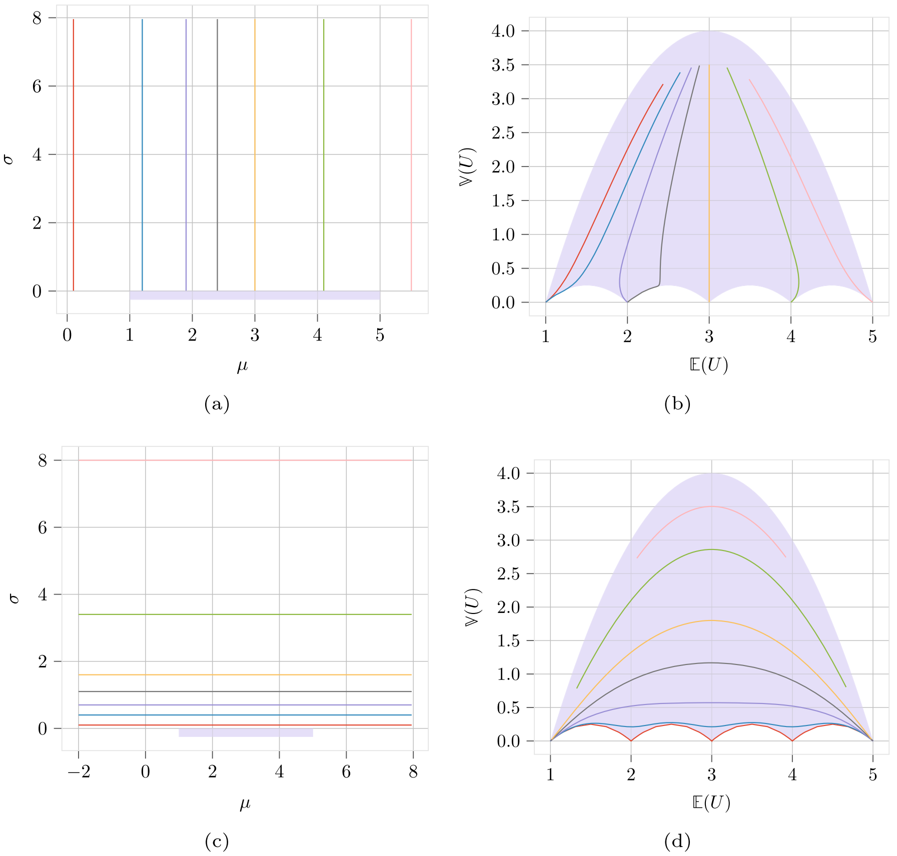
(d)

**Fig. 5** GSD distributions of $U$ for $M = 5$ and for various values of $\psi$ and $\rho$

## 2.3 Comparing ordered probit's and GSD's parametrisation

In this section, we present the interaction between ordered probit parameters, GSD parameters, and the $(\mathbb{E}(U), \mathbb{V}(U))$ space. Figure 6 presents the interaction between ordered probit parameters ($\mu$ and $\sigma$) and summary statistics ($\mathbb{E}(U)$ and $\mathbb{V}(U)$). The latter are calculated taking discrete responses generated by the ordered probit model with a given $\mu$ and $\sigma$ pair. The lines present in the left-hand-side of Fig. 6, correspond to the same coloured lines in the right-hand-side of the same figure. Specifically, the ordering of lines (when going from left to right in Fig. 6a and top to bottom in Fig. 6c) in the left-hand-side of the figure is the same as the ordering of lines in the right-hand-side of the figure. Figure 7 presents the corresponding plots for the GSD class. Note, however, that the ordering of lines in Fig. 7c is reversed to the ordering of lines in Fig. 7d. Differently put, the top-most line in Fig. 7c corresponds to the bottom-most line in Fig. 7d. Importantly, both Figs. 6 and 7 use $M = 5$.

Figure 6 is a graphical presentation of the problems inherent to the estimation and interpretation of ordered probit parameters, when these are based on the observations of the random variable $U$. First of all, the mapping from any bounded set of $(\mu, \sigma)$ pairs, results in a set of $(\mathbb{E}(U), \mathbb{V}(U))$ pairs that does not cover the whole ghost-like area of



(a)

(b)

(c)

(d)

**Fig. 6** Mapping of ordered probit parameters to the $(\mathbb{E}(U), \mathbb{V}(U))$ space. The violet ghost-like area marks all possible $(\mathbb{E}(U), \mathbb{V}(U))$ pairs for a discrete process with values $\{1, 2, 3, 4, 5\}$

all possible $(\mathbb{E}(U), \mathbb{V}(U))$ pairs. Notice that the lines in Fig. 6b do not reach neither the sides nor the top of the ghost-like area. Second of all, there is no analytical formula mapping $(\mathbb{E}(U), \mathbb{V}(U))$ pairs to $(\mu, \sigma)$ pairs. It is difficult to even approximately guess which $(\mu, \sigma)$ pair corresponds to which $(\mathbb{E}(U), \mathbb{V}(U))$ pair. This is a big limitation of ordered probit's parameterisation. The estimation of a $(\mathbb{E}(U), \mathbb{V}(U))$ pair, based on observations $U_1, \ldots, U_n$, is an easy task. Unfortunately, the results of this estimation are not useful for estimating ordered probit parameters.

The problems described above do not apply to GSD's parameterisation. The $\psi$ parameter is the expected value of the observations $(U_1, \ldots, U_n)$. Thus, for any $(\mathbb{E}(U), \mathbb{V}(U))$ pair, we immediately know the corresponding $\psi$. This is because $\psi$ is equal to $\mathbb{E}(U)$. Notice that the vertical lines in Fig. 7a and b are in identical positions along the horizontal axis. The second GSD class' parameter, $\rho$, is the confidence parameter. Its value of 0 corresponds to the biggest possible variance (the upper bound of the violet coloured ghost-like area in Fig. 7b) and 1 corresponds to the smallest possible variance (the lower bound of the violet coloured ghost-like area

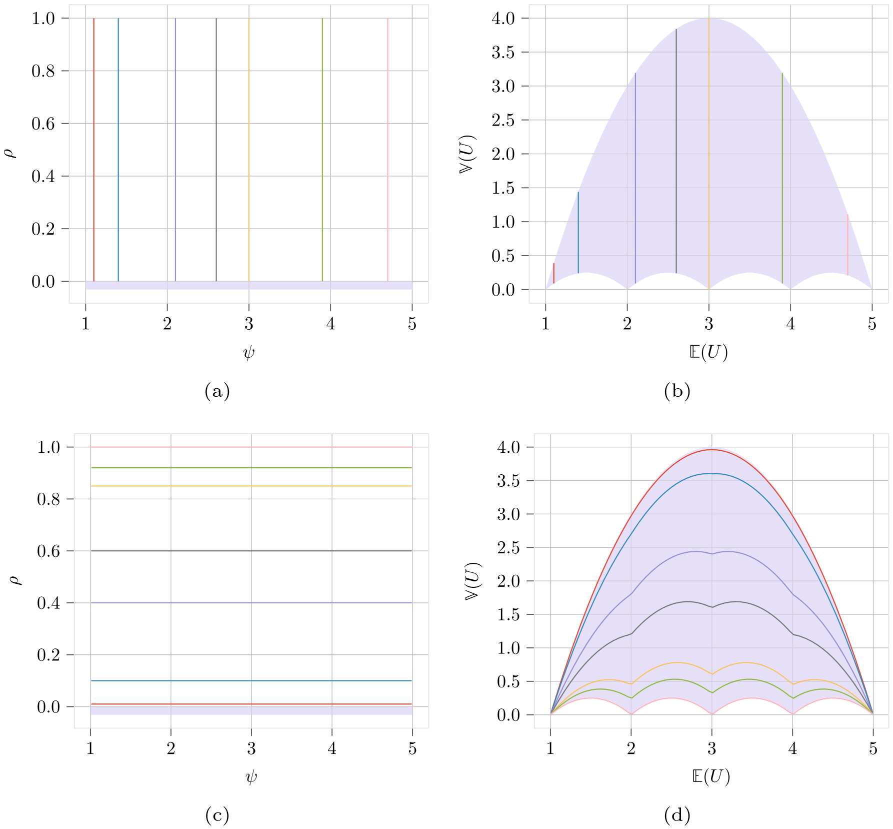


**Fig. 7** Mapping of GSD parameters to the $(\mathbb{E}(U), \mathbb{V}(U))$ space. The violet ghost-like area marks all possible $(\mathbb{E}(U), \mathbb{V}(U))$ pairs for a discrete process with values $\{1, 2, 3, 4, 5\}$

in Fig. 7b) of the observations. Moreover, the range of $\rho$ values is linear. For example, $\rho = 0.7$ can be interpreted as 70%, in terms of the available variance. That is, $\mathbb{V}(U) = 70\% V_{\min}(\psi) + 30\% V_{\max}(\psi)$. Therefore, to obtain $\rho$ from a $(\mathbb{E}(U), \mathbb{V}(U))$ pair, it is enough to calculate the distance between $\mathbb{V}(U)$ and the upper bound of the violet coloured ghost-like area in Fig. 7b. This distance should be then divided by the distance between the upper and lower bounds of the violet coloured ghost-like area in Fig. 7b. Differently put, $\rho = \frac{V_{\max}(\psi)-\mathbb{V}(U)}{V_{\max}(\psi)-V_{\min}(\psi)}$. Value of $\rho$ is thus simply a ratio between the distance between the observed variance ($\mathbb{V}(U)$) and the maximum possible variance ($V_{\max}$) and the available range of variance ($V_{\max}(\psi) - V_{\min}(\psi)$). The easy mapping of $(\mathbb{E}(U), \mathbb{V}(U))$ pairs to $(\psi, \rho)$ pairs makes the estimation of $(\psi, \rho)$ pairs (when based on $U_1, \ldots, U_n$ observations) much easier than the estimation of $(\mu, \sigma)$ pairs for the ordered probit model.

We would also like to draw the reader's attention to the violet horizontal bars present in Figs. 6a, c, and 7a, c. In all cases, the violet bars span the range from 1 to 5. This range corresponds to the width of the violet ghost-like area in the right-hand side of Figs. 6 and 7. Please note that GSD's parameterisation is bounded to this 1 to 5 range, whereas ordered probit's one is not. This is yet another advantageous feature of GSD's parameterisation. Being bounded to the same range as the range of observable responses, GSD class parameters are easier to interpret and understand.

## 3 GSD parameters estimation

Ordered probit and GSD models cannot be used without an accurate and efficient parameter estimation procedure. The simplest approach is to use the method of moments. As we mentioned in the previous section, the method of moments for the ordered probit model is rather problematic. However, for the GSD class, moments based estimation is quite simple, i.e.,

$$(\hat{\psi}, \hat{\rho}) = \left( \widehat{\mathbb{E}(U)}, \frac{V_{\max}(\widehat{\mathbb{E}(U)}) - \widehat{\mathbb{V}(U)}}{V_{\max}(\widehat{\mathbb{E}(U)}) - V_{\min}(\widehat{\mathbb{E}(U)})} \right),$$

where $\widehat{\mathbb{E}(U)}$, $\widehat{\mathbb{V}(U)}$ are expectation value and variance of the empirical distribution. To compare the GSD with ordered probit and to apply the likelihood ratio test of goodness-of-fit, we actually use Maximum Likelihood Estimator (MLE) for both models. To make the comparison between the two models fair, when fitting them to real data, we use the same numerical estimation method for both, i.e., we use the estimation using a dense grid of points. Specifically, we first compute probabilities of all response categories for a set of $(\psi, \rho)$ (or $(\mu, \sigma)$) pairs and then search through the resultant grid to find the pair best matching the sample of interest. In the multidimensional case (cf. Appendix E), for generated data, we use the gradient based estimation method for the GSD class. The moments based estimator serves as a starting point. The exact formulae for the log-likelihood function and gradient that were used in the estimation algorithm are in Appendix B.

### 3.1 Numerical experiments for the GSD class

To validate our MLE procedure, we perform a simulation study. We draw data from the proposed distribution and then estimate the distribution parameters using the samples generated. Importantly, we do so for the case of $M = 5$.

In Fig. 8 we present the risk measured as Root Mean Square Distance (RMSD) between true value $\psi$ and estimated $\hat{\psi}$, for sample sizes $n = 12, 24, 50, 200$. As one can see, the hardest case is when $\rho$ is small and $\psi$ is in the middle of the $[1, M]$ scale. This behaviour is expected, since in the middle of the scale, the possible variance is the largest (cf. Fig. 6b). One can also see that the parameter $\psi$ is rather easy to estimate. That is, the risk is rather small for $\psi \in [1, 5]$, even for small sample sizes.

In Fig. 9 we present the risk measured as the RMSD between true value $\rho$ and estimated $\hat{\rho}$, for sample sizes $n = 12, 24, 50, 200$. In this case, the situation is slightly more complicated (than that presented above for $\psi$). The hardest case is when $\psi$ is on the edge of the $[1, M]$ interval. As one can see in Fig. 4, when $\psi$ is close to the edge of the $[1, M]$ interval, variance expressed as a function of $\rho$ is almost horizontal. This

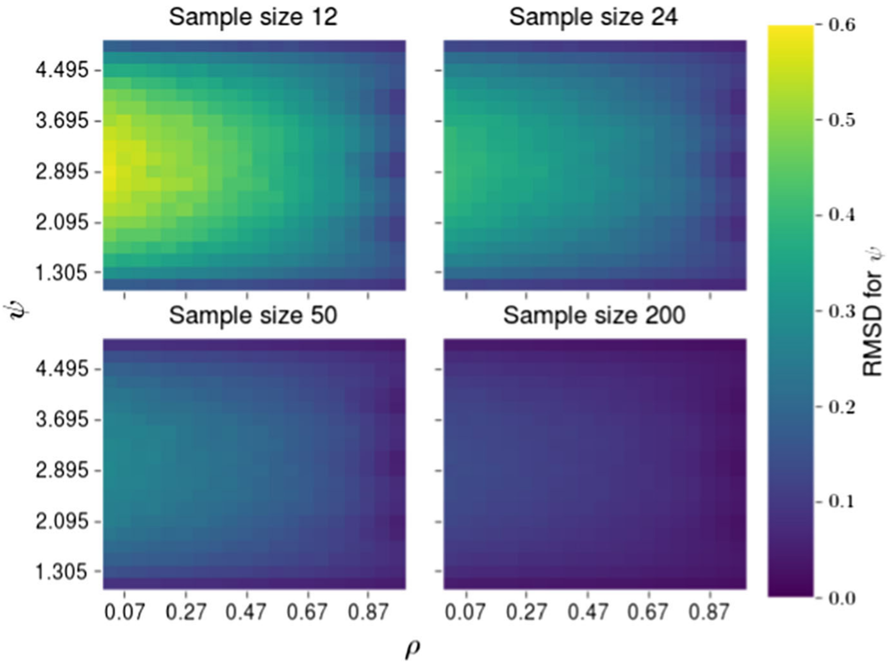


**Fig. 8** Root mean square distance (RMSD) between the input $\psi$ and the estimated $\hat{\psi}$ for different $(\psi, \rho)$, different sample sizes, and $M = 5$. For every sample size we generated 500,000 samples to obtain the figure

means that even small changes in sample variance correspond to large changes of $\rho$. Simply put, the smaller the $[V_{\min}(\psi), V_{\max}(\psi)]$ interval, the harder the estimation of $\rho$. It is worth pointing out, however, that even for hard cases the risk is getting smaller, the larger is the sample size.

We also validate the MLE procedure for the multidimensional case. Specifically, we generate responses of $n$ raters (also referred to as subjects), each having assigned a confidence parameter $\rho_1, \ldots, \rho_n$. The subjects rate $m$ objects (also referred to as stimuli), each having assigned an expectation value $\psi_1, \ldots, \psi_m$ (which can be, for example, interpreted as latent true qualities of a set of $m$ videos presented to $n$ subjects). Based on the responses generated, we used the multidimensional MLE to recover the GSD parameters. More details and estimation results are in Appendix E.

## 4 Real data example for multimedia quality assessment

In the previous sections, we presented a new GSD family of distributions and showed that the estimation method properly extracts GSD parameters when applied to the simulated data. In this section, we validate if the proposed GSD class can be used to model real subjective data (i.e., subjective responses coming from MQA experiments).

In MQA experiments, multiple participants assess the quality of multiple stimuli. To be more precise, there are usually around 24 participants, who assess the quality

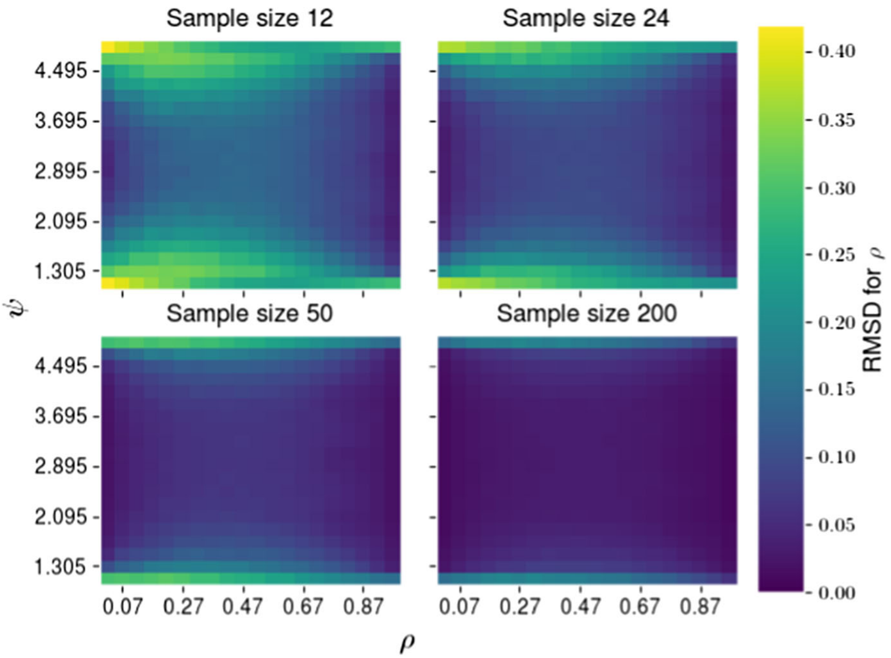


**Fig. 9** Root mean square distance (RMSD) between the input $\rho$ and the estimated $\hat{\rho}$ for different $(\psi, \rho)$, different sample sizes, and $M = 5$. For every sample size we generated 500,000 samples to obtain the figure

of roughly 160 stimuli, using the five-level assessment scale (cf. Sect. 2.2.1). In the great majority of experiments, each participant assesses the quality of each stimulus exactly once. Thanks to this, and since our focus in this paper is on the per stimulus analysis, we treat responses as independent observations of a random variable of interest. MQA experiments measure *subjectively perceived* quality. Hence, there is no perfect agreement between participants. Even the same person assessing the same stimulus multiple times tends to assign to it different response categories (Perez et al. 2021).

### 4.1 Comparing goodness-of-fit of ordered probit and GSD

We want to compare the GSD with the ordered probit model and one state-of-the-art solution. We come up with the latter by adapting the model presented in Li et al. (2020). We call the adapted version of the model *Simplified Li2020* or SLI for short.[3]

To check whether a distribution fits specific data, we have to perform a two-step procedure. The first step is to estimate distribution parameters for a sample of interest. The second step is to test a null hypothesis, saying that the sample truly comes from the assumed distribution (GSD, ordered probit or SLI), given the parameters estimated in the first step. We choose a standard likelihood ratio approach to test the goodness-of-fit

[3] For a detailed description regarding how did we transform the model from Li et al. (2020), we refer the reader to Sect. III-D of Nawała et al. (2022).

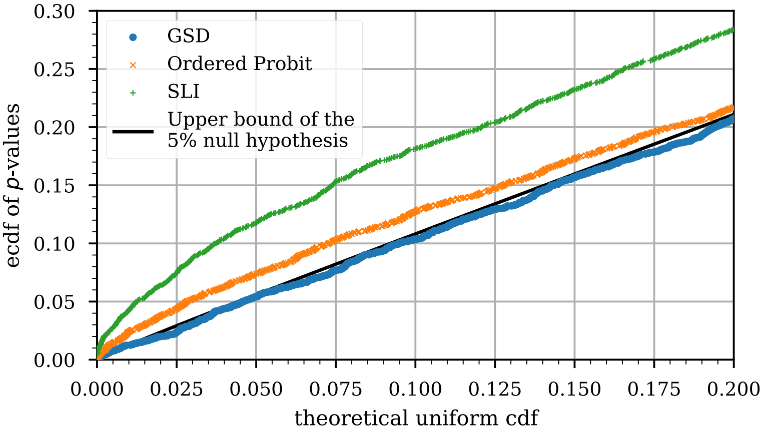


**Fig. 10** $p$-Value P–P plot for typical multimedia quality assessment (MQA) experiments. $p$-Values come from the G-test of goodness-of-fit applied to the GSD, ordered probit and Simplified Li2020 (SLI) models fitted to responses from 24 real-life subjective experiments. Ecdf stands for empirical cumulative distribution function

(GoF) of the models. Specifically, we use the G-test of GoF (cf. Sect. 14.3.4 of Agresti (2002)). Since sample sizes we consider are mostly small (less than 30 observations per sample), we do not use the asymptotic distribution for calculating the $p$-value. Conversely, we estimate the $p$-value using a bootstrapped version of the G-test (see Appendix C). (For comprehensive theoretical considerations on the topic please take a look at Efron and Tibshirani (1993).)

The data we use to perform the analysis come from six MQA studies: (i) ITU (TU-T Study Group 12 1998), (ii) HDTV (Pinson et al. 2010), (iii) MM2 (Pinson et al. 2012), (iv) 14-505 (Pinson and Janowski 2014), (v) ITS4S (Pinson 2018) and (vi) NFLX. We do not provide here extensive details regarding each study. Instead, we refer the reader to publications cited next to each acronym. Since the NFLX study does not have a dedicated publication, we refer the reader to Sect. II.C of Nawała et al. (2022). Two important features of the six studies is that they all focus on Multimedia Quality Assessment (MQA) and follow best practices and recommendations in the field. In other words, we can safely call them *typical* MQA studies. Some studies represent data from more than one subjective experiment. Differently put, one study may consist of multiple experiments. In total, we use data from 24 subjective experiments. This amounts to more than 100,000 individual scores (exactly 111,198) for more than 3500 stimuli (exactly 3643).

In Fig. 10 we present the cumulative distribution function (CDF) of GoF test $p$-values for the GSD, ordered probit, and SLI models. The black line is the upper bound of 95% right-sided confidence interval for the CDF of $p$-values under the null hypothesis. Specifically, under the null hypothesis, the CDF of $p$-values is not greater than the uniform distribution function (for more details see Nawała et al. (2020)). As one can see, there is no evidence that the GSD is not the correct way of modelling subjective responses from MQA experiments. On the other hand, there is evidence that the distributions modelled by the ordered probit and SLI models are not suitable here.

### 4.2 Bootstrapping

If we would like to use the bootstrap technique in some testing problems with data expressed on a Likert scale, there are two approaches to resampling that one can consider. One can either use the empirical distribution or fit a distribution coming from some assumed parametric class. Here, we show that for at least one type of real data, specifically responses coming from MQA experiments, it is better to use the estimated GSD than it is to use the empirical distribution. This holds at least in the case of relatively small sample sizes. (In the field of MQA, usually only up to 30 responses per stimulus are available.) To compare the behaviour of empirical probability mass function (EPMF) and the GSD, we use the following algorithm.

Let us denote by $N$ the number of observations in the large sample (e.g., $N = 200$) and by $n$ the number of observations in the subsample of this large sample (e.g., $n = 24$). Now, we denote by $(N_1, N_2, N_3, N_4, N_5)$ the frequencies of each response category in the large sample. We denote by $(p_1, p_2, p_3, p_4, p_5)$ the EPMF of the large sample. The test procedure is as follows (assuming there are five response categories):

1. Generate $MC$ bootstrap samples (e.g., $MC = 10{,}000$) of size $n$ from the EPMF of the large sample $(p_1, p_2, p_3, p_4, p_5)$.
2. For the $r$-th bootstrap sample ($r = 1, 2, \ldots, MC$) do the following.
   (a) Estimate response category probabilities using maximum likelihood estimation for the model of interest (e.g., the GSD model). Denote the estimated probabilities by $\big(\hat{q}_1, \hat{q}_2, \hat{q}_3, \hat{q}_4, \hat{q}_5\big)$.
   (b) Denote by $\big(\hat{v}_1, \hat{v}_2, \hat{v}_3, \hat{v}_4, \hat{v}_5\big)$ the EPMF of the bootstrap sample.
   (c) Find the likelihood $\mathcal{L}_m$ of the estimated model for the large sample. In other words, calculate

$$\mathcal{L}_m = \prod_{k=1, N_k \neq 0}^{5} \hat{q}_k^{N_k}.$$

   (d) Find the likelihood $\mathcal{L}_e$ of the bootstrap sample's EPMF for the large sample. In other words, calculate

$$\mathcal{L}_e = \prod_{k=1, N_k \neq 0}^{5} \hat{v}_k^{N_k}.$$

   (e) Find the natural logarithm of the ratio of the two likelihoods and denote it by $W_r$

$$W_r = \ln\left(\frac{\mathcal{L}_m}{\mathcal{L}_e}\right).$$

Note that the above simplifies to

$$W_r = \sum_{k=1, N_k \neq 0}^{5} N_k \left( \ln \hat{q}_k - \ln \hat{v}_k \right).$$

3. Calculate the estimator of $p_{\text{GSD}} - p_{\text{e}} = P(W_r > 0) - P(W_r < 0)$, which is the difference between the probability that the GSD has greater likelihood than the EPMF and the probability that the EPMF has greater likelihood than the GSD. This can be formally described by the following.

$$\hat{p}_{\text{GSD}} - \hat{p}_{\text{e}} = \frac{\sum_{r=1}^{MC} I\,(W_r > 0)}{MC} - \frac{\sum_{r=1}^{MC} I\,(W_r < 0)}{MC},$$

where $I(x)$ is one if $x$ is true or 0 if $x$ is false.

4. Calculate .95 confidence interval for $p_{\text{GSD}} - p_{\text{e}}$ i.e.,

$$L = \hat{p}_{\text{GSD}} - \hat{p}_{\text{e}} - 1.96\sqrt{\frac{\hat{p}_{\text{GSD}} + \hat{p}_{\text{e}} - (\hat{p}_{\text{GSD}} - \hat{p}_{\text{e}})^2}{MC}}$$

$$R = \hat{p}_{\text{GSD}} - \hat{p}_{\text{e}} + 1.96\sqrt{\frac{\hat{p}_{\text{GSD}} + \hat{p}_{\text{e}} - (\hat{p}_{\text{GSD}} - \hat{p}_{\text{e}})^2}{MC}}$$

For $L > 0$ the GSD performs better. For $R < 0$ the EPMF performs better. If $[L, R]$ contains zero there is no significant difference between the GSD and EPMF.

We use data from four MQA studies: (i) MM2 (Pinson et al. 2012), (ii) HDTV (Pinson et al. 2010), (iii) NFLX (cf. Sect. II.C of Nawała et al. (2022)) and (iv) ITERO (Perez et al. 2021). We describe the first three studies in Sect. 4.1. The last study, contrary to the first three, is not a typical MQA study. We decide to use it anyway for two reasons. First, being atypical, it should not not give unfair advantage to the GSD. Second, it provides real data with many responses per stimulus. This last property also stands behind our choice to use the other three studies (i.e., MM2, HDTV and NFLX). Specifically, we only select from these stimuli with at least 144 responses. This results in 234 stimuli, each assigned between 144 and 228 responses.

We use three small sample sizes, i.e., $n = \{12, 24, 50\}$. This way we can observe how the GSD performs (when compared to the empirical distribution) for different fractions of the large sample information available. Intuitively, we expect the empirical distribution's performance to improve as the small sample size increases. If the GSD proofs to perform differently than the empirical distribution we would observe how the increasing small sample size influences the difference between the two approaches. We emphasise here that the increasing small sample size always favours the empirical distribution. On the other hand, the performance of the GSD depends on how well it fits to the distribution of responses observed in the large sample. If the fit is good, increasing small sample size also favours the GSD. If the fit is poor, increasing small sample size does not necessarily improve GSD's performance.

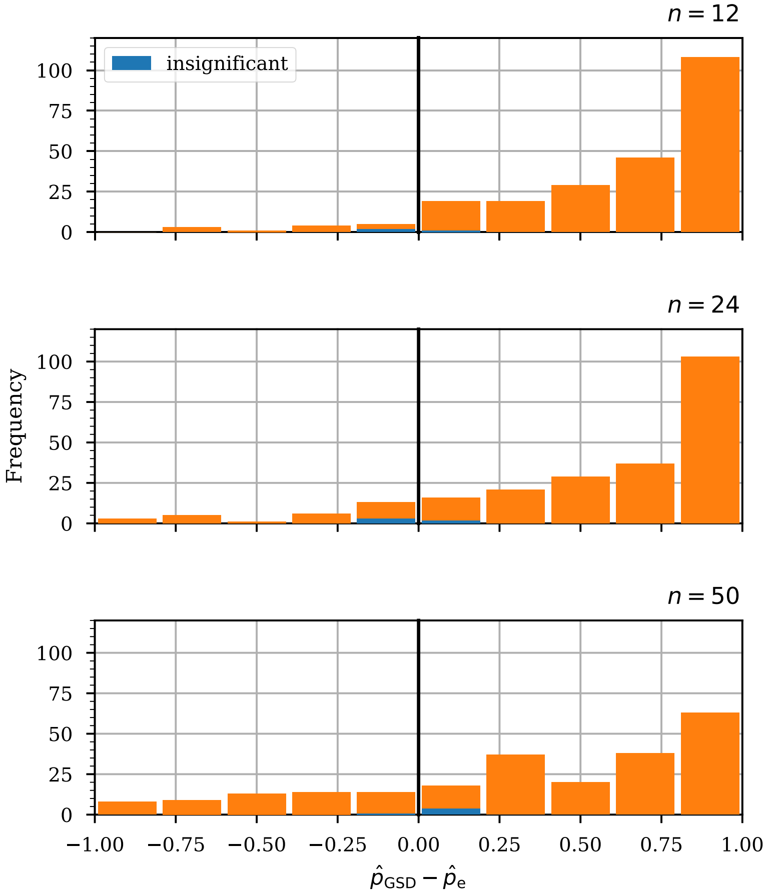


**Fig. 11** Histograms depicting the distribution of probability differences $\hat{p}_{\mathrm{GSD}} - \hat{p}_{\mathrm{e}}$ for three different small sample sizes (i.e., 12, 24 and 50) and for the case of the (unmodified) GSD being compared with the (unmodified) empirical distribution. Blue-coloured parts of the bars represent statistically insignificant probability differences

Figure 11 presents results of the analysis. It contains three histograms of probability differences $\hat{p}_{\mathrm{GSD}} - \hat{p}_{\mathrm{e}}$. Each histogram shows results for one of the investigated small sample sizes (i.e., 12, 24, and 50). Larger probability mass to the right of zero means the GSD outperforms the empirical distribution. Larger probability mass to the left of zero means the empirical distribution performs better than the GSD. As can be seen, the GSD outperforms the empirical distribution for all three small sample sizes considered.

We theorise that the reason the GSD outperforms the empirical distribution is because the latter assigns a larger than the GSD probability to empty cells (i.e., response categories with no responses assigned to them in the sample of interest). To verify this claim, we run our analysis once again, this time modifying the estimation procedure both for the GSD and empirical distribution. This modification does not allow any

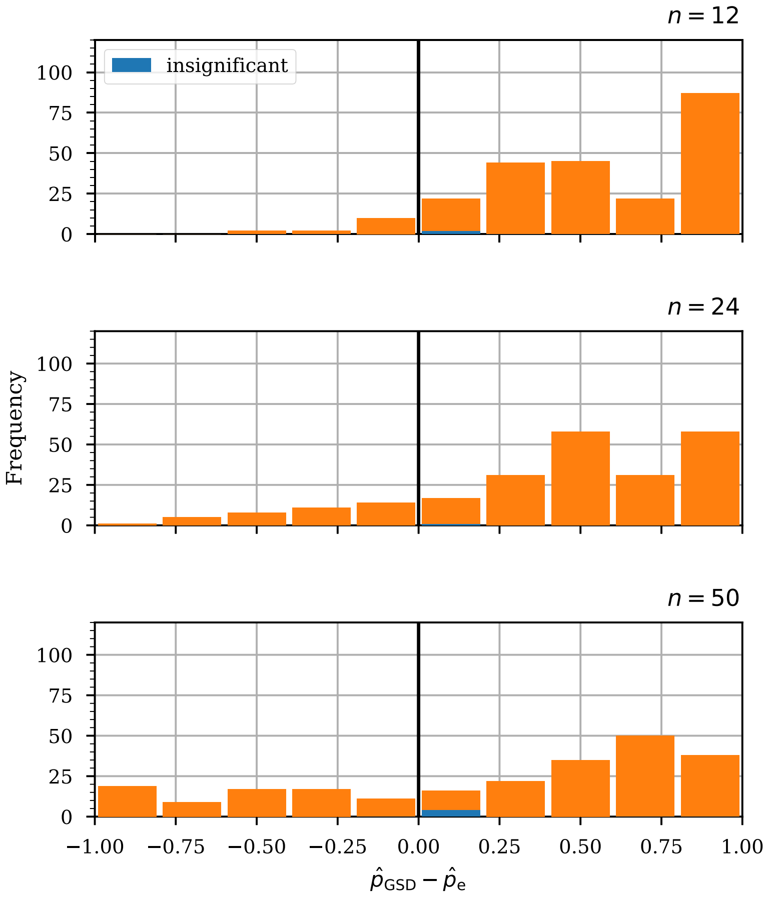


**Fig. 12** Histograms depicting the distribution of probability differences $\hat{p}_{\mathrm{GSD}} - \hat{p}_{\mathrm{e}}$ for three different small sample sizes (i.e., 12, 24 and 50) and for the case of the corrected GSD being compared with the corrected empirical distribution. Blue-coloured parts of the bars represent statistically insignificant probability differences

empty cells. In other words, the estimated probability of any response category has to be necessarily in the interval $(0, 1)$. The details are in Appendix D. Figure 12 presents results for the case of the corrected GSD being compared with the corrected empirical distribution. Again, the GSD outperforms the empirical distribution, although this time by a smaller margin.

The results clearly show that the GSD is a better choice than the empirical distribution when it comes to resampling of subjective responses from MQA studies.

## 5 Conclusion

In this paper, we propose a Generalised Score Distribution (GSD) class. It is a family of discrete distributions with: finite support, two parameters, and no more than one change in probability monotonicity. The distribution parameters are: $\psi$ determining the mean, and $\rho$ determining the spread of the responses.

We show the usefulness of the GSD class for modelling, with a special focus on the Multimedia Quality Assessment (MQA) field. The class is a convenient regularisation of the multinomial distribution. The GSD class has only two parameters and covers all possible first- and second-order moments for a distribution defined on a discrete finite support. We also evidence that the GSD class can be useful in testing problems using the parametric bootstrap technique.

The advantage of the GSD class is that its $\rho$ parameter can be used to determine the type of the underlining process. With $\rho$ close to 1, we know that the process is similar to the Bernoulli distribution. Likewise, for $\rho < C(\psi)$ we know the process rather resembles the beta-binomial distribution. This information can be used as a diagnostic tool, answering the following question: "What is the spread of the responses?" Note that the GSD class can be easily used outside of the MQA field, wherever information about responses spread is relevant.

We strongly believe that the GSD class can be of use for modelling results similar in nature to those reported in the field of MQA. More specifically, the GSD can be potentially useful for modelling subjective responses, where the population of observers generally agree about a given trait of a stimulus presented to them (e.g., about the visual quality of a distorted image). To put this differently, the GSD will not likely work in cases where there are evident subgroups of observers. For example, when there are two groups that have opposing views on a stimulus trait, they are asked to assess. The only exception to GSD's inability of modelling opposing views is the so-called "love or hate" case. In this case, a significant proportion of the population of observers either scores a stimulus trait extremely high or extremely low (cf. Fig. 5, the GSD distribution with $\psi = 2.85$ and $\rho = 0.38$).

In the future research, we would like to add more parameters to the GSD class. One idea is to add a parameter related to a potential personal bias of each observer (similarly to what is done in Janowski and Pinson (2015) and Li and Bampis (2017)). This subject bias parameter may, for example, show whether a person is generally more optimistic than other raters are. Another interesting direction of research would be using the GSD for data other than that coming from MQA subjective experiments. We would like to collaborate with scientists working in different fields, from audio and image quality, through student performance assessment, and up to psychology and sociology. In all those fields, a proper modelling of the response generation process would help to gain new insights. Our results obtained for MQA subjective data might be treated as a proof of concept, showing that the GSD class may be of use for those other fields as well.

**Acknowledgements** The authors would like to thank Netflix, Inc. for sponsoring the initial phases of this research. This work was also partially supported by the PL-Grid Infrastructure. The work of Jakub Nawała and Lucjan Janowski have received funding from the Norwegian Financial Mechanism 2014-2021 under project 2019/34/H/ST6/00599. We warmly acknowledge comments and suggestions from Zhi Li of Netflix as well. We would also like to thank Anush Krishna Moorthy for coordinating the NFLX experiment and providing valuable feedback. Furthermore, we would like to thank Pablo Pérez, Narciso García and Margaret Pinson, for creating the ITERO data set, which helped us perform the bootstrap analysis (cf. Sect. 4.2).

## Declarations

**Conflict of interest** The authors have no competing interests to declare that are relevant to the content of this article.

## Appendix A: Proofs

***Proof of Proposition 1*** First notice that $P_{G_\rho}(\epsilon = k - \psi)$ (see formula (11)) can be rewritten as

$$P_{G_\rho}(\epsilon = k - \psi) = \binom{M-1}{k-1} \frac{\mathcal{B}\left(\frac{(\psi-1)\rho}{(M-1)(C(\psi)-\rho)} + k - 1, \frac{(M-\psi)\rho}{(M-1)(C(\psi)-\rho)} + M - k\right)}{\mathcal{B}\left(\frac{(\psi-1)\rho}{(M-1)(C(\psi)-\rho)}, \frac{(M-\psi)\rho}{(M-1)(C(\psi)-\rho)}\right)}$$

Now observe that $\epsilon + \psi - 1$ has the beta-binomial distribution $BB(M-1, \alpha, \beta)$ with parameters

$$\alpha = \frac{(\psi - 1)\rho}{(M-1)(C(\psi) - \rho)}, \quad \beta = \frac{(M - \psi)\rho}{(M-1)(C(\psi) - \rho)}.$$

Using formula for beta-binomial expectation value we obtain

$$\mathbb{E}(U) = \mathbb{E}(\psi + \epsilon) = \frac{(M-1)\alpha}{\alpha + \beta} + 1 = \psi.$$

Using formula for beta-binomial variance we obtain

$$\mathbb{V}(U) = \mathbb{V}(\psi + \epsilon) = \mathbb{V}(\psi + \epsilon - 1) = \frac{(M-1)\alpha\beta(\alpha + \beta + M - 1)}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$
$$= \frac{(\psi - 1)(M - \psi)}{M - 1}\left(1 + \frac{M-2}{C(\psi)}(C(\psi) - \rho)\right).$$

Since $(\psi - 1)(M - \psi) = V_{\max}(\psi)$ and $C(\psi) = \frac{M-2}{M-1} \frac{V_{\max}(\psi)}{V_{\max}(\psi) - V_{\min}(\psi)}$ (see formulas (6), (7) and (10)) we have

$$\begin{aligned}\mathbb{V}(U) &= \frac{V_{\max}(\psi)}{M-1} + (V_{\max}(\psi) - V_{\min}(\psi)) \left( \frac{M-2}{M-1} \frac{V_{\max}(\psi)}{V_{\max}(\psi) - V_{\min}(\psi)} - \rho \right) \\ &= \rho V_{\min}(\psi) + (1-\rho) V_{\max}(\psi).\end{aligned}$$

□

***Proof of Proposition 2*** In $F_\rho$ distribution case (see formula (13)) notice that the random variable $\epsilon$ is a mixture of random variables $\epsilon_1$ and $\epsilon_2$, i.e.,

$$\epsilon = D\epsilon_1 + (1-D)\epsilon_2,$$

where

$$\begin{aligned}&P(D=1) = \frac{\rho - C(\psi)}{1 - C(\psi)}, \quad P(D=0) = \frac{1-\rho}{1-C(\psi)}, \\ &P(\epsilon_1 = k - \psi) = [1 - |k-\psi|]_+, \\ &P(\epsilon_2 = k - \psi) = \binom{M-1}{k-1} \left( \frac{\psi - 1}{M-1} \right)^{k-1} \left( \frac{M-\psi}{M-1} \right)^{M-k},\end{aligned}$$

and $\epsilon_1, \epsilon_2, D$ are independent. If $\psi$ is an integer $P(\epsilon_1 = 0) = 1$ so $\mathbb{E}(\epsilon_1) = 0$. In case $\psi$ is not an integer we have

$$\mathbb{E}(\epsilon_1) = (1 + \lfloor \psi \rfloor - \psi)(\lfloor \psi \rfloor - \psi) + (\psi - \lfloor \psi \rfloor)(1 + \lfloor \psi \rfloor - \psi) = 0.$$

In both cases $\mathbb{E}(\epsilon_1) = 0$. Now, observe that $\epsilon_2 + \psi - 1$ has the binomial distribution $B(M-1, p)$ with $p = \frac{\psi - 1}{M-1}$. Therefore

$$\mathbb{E}(\epsilon_2) = (M-1)\frac{\psi - 1}{M-1} - \psi + 1 = 0.$$

We have then

$$\begin{aligned}\mathbb{E}(U) = \mathbb{E}(\epsilon + \psi) = \mathbb{E}(\epsilon_1 \\ + \psi | D = 1) P(D=1) + \mathbb{E}(\epsilon_2 + \psi | D = 0) P(D = 0) = \psi\end{aligned}$$

Now, notice that $\mathbb{V}(\epsilon_1) = V_{\min}(\psi)$ and $V(\epsilon_2) = V_{\mathrm{Bin}}(\psi)$ (see formulas (6), (7) and (8)). Therefore

$$\begin{aligned}\mathbb{V}(U) = \mathbb{V}(\epsilon) = \mathbb{E}(\epsilon^2) &= \mathbb{E}(\epsilon_1^2)\frac{\rho - C(\psi)}{1 - C(\psi)} + \mathbb{E}(\epsilon_2^2)\frac{1-\rho}{1-C(\psi)} \\ &= V_{\min}(\psi)\frac{\rho - C(\psi)}{1 - C(\psi)} + V_{\mathrm{Bin}}(\psi)\frac{1-\rho}{1-C(\psi)} =: f_\psi(\rho).\end{aligned}$$

We want to show that for every fixed $\psi \in [1, M]$, the variance $\mathbb{V}(U)$ is the linear function of $\rho$ equal to $\rho V_{\min}(\psi) + (1-\rho)V_{\max}(\psi)$. Notice that the $f_\psi$ function is a linear function of variable $\rho$. It is easy to see that $f_\psi(1) = V_{\min}(\psi)$ so let us check the derivative. Since $C(\psi) = \frac{M-2}{M-1}\frac{V_{\max}(\psi)}{V_{\max}(\psi)-V_{\min}(\psi)}$ and $V_{\mathrm{Bin}}(\psi) = \frac{V_{\max}(\psi)}{M-1}$, we obtain

$$\begin{aligned}\frac{d}{d\rho}f_\psi(\rho) &= \frac{(V_{\max}(\psi)-V_{\min}(\psi))((M-1)V_{\min}(\psi)-V_{\max}(\psi))}{(M-1)(V_{\max}(\psi)-V_{\min}(\psi))-(M-2)V_{\max}(\psi)}\\ &= V_{\min}(\psi)-V_{\max}(\psi) = \frac{d}{d\rho}(\rho V_{\min}(\psi)+(1-\rho)V_{\max}(\psi))\end{aligned}$$

Therefore

$$\mathbb{V}(U) = \rho V_{\min}(\psi) + (1-\rho)V_{\max}(\psi).$$

□

***Proof of Proposition 4*** First notice that in the case of $\rho \geq C(\psi)$

$$P(Z_i = 1) = \begin{cases} \frac{\rho-C(\psi)}{1-C(\psi)} + \frac{1-\rho}{1-C(\psi)}\frac{\psi-1}{M-1} \ \text{for} \ i \leq \lfloor\psi\rfloor - 1 \\ \frac{\rho-C(\psi)}{1-C(\psi)}(\psi+1-\lceil\psi\rceil) + \frac{1-\rho}{1-C(\psi)}\frac{\psi-1}{M-1} \ \text{for} \ i = \lceil\psi\rceil - 1 \\ \frac{1-\rho}{1-C(\psi)}\frac{\psi-1}{M-1} \ \text{for} \ i \geq \lceil\psi\rceil \end{cases}.$$

Random variables $Z_i$ can be written as $Z_i = DX_i + (1-D)Y_i$ where $X_i, Y_i, D$ are independent zero–one random variables and

$$\begin{aligned} P(D=1) &= \frac{\rho - C(\psi)}{1-C(\psi)}\\ P(X_i = 1) &= \begin{cases} 1 \ \text{for} \ i \leq \lfloor\psi\rfloor - 1 \\ \psi + 1 - \lceil\psi\rceil \ \text{for} \ i = \lceil\psi\rceil - 1 \\ 0 \ \text{for} \ i \geq \lceil\psi\rceil \end{cases}\\ P(Y_i = 1) &= \frac{\psi - 1}{M-1}. \end{aligned}$$

Now, observe that

$$P\left(\sum_{i=1}^{M-1} X_i = k-1\right) = [1 - |k-\psi|]_+$$

and

$$P\left(\sum_{i=1}^{M-1} Y_i = k-1\right) = \binom{M-1}{k-1}\left(\frac{\psi-1}{M-1}\right)^{k-1}\left(\frac{M-\psi}{M-1}\right)^{M-k}$$

for $k = 1, \dots, M$. Since

$$U - 1 = \sum_{i=1}^{M-1} Z_i = \sum_{i=1}^{M-1} (DX_i + (1-D)Y_i) = D \sum_{i=1}^{M-1} X_i + (1-D) \sum_{i=1}^{M-1} Y_i$$

then

$$\begin{aligned} &P(U = k) \\ &= \frac{\rho - C(\psi)}{1 - C(\psi)}[1 - |k - \psi|]_+ + \frac{1-\rho}{1-C(\psi)} \binom{M-1}{k-1} \left(\frac{\psi - 1}{M-1}\right)^{k-1} \left(\frac{M - \psi}{M-1}\right)^{M-k} \end{aligned}$$

for $k = 1, \dots, M$. The case $\rho < C(\psi)$ is an easy consequence of the fact that $U - 1$ has the beta-binomial distribution $BB(M-1, \alpha, \beta)$ with parameters

$$\alpha = \frac{(\psi - 1)\rho}{(M-1)(C(\psi) - \rho)}, \quad \beta = \frac{(M - \psi)\rho}{(M-1)(C(\psi) - \rho)}.$$

□

## Appendix B: Formula for the gradient of GSD's log-likelihood function

Denote by $(n_1, \dots, n_M)$ numbers of observed responses and

$$\begin{aligned} V'_{\min}(\psi) &= -2\psi + \lceil \psi \rceil + \lfloor \psi \rfloor, \\ V'_{\max}(\psi) &= -2\psi + M + 1, \\ C'(\psi) &:= \frac{M-2}{M-1} \frac{V_{\max}(\psi)V'_{\min}(\psi) - V'_{\max}(\psi)V_{\min}(\psi)}{(V_{\max}(\psi) - V_{\min}(\psi))^2}. \end{aligned}$$

The Log-Likelihood function for $\rho < C(\psi)$ is equal to

$$\begin{aligned} l(\psi, \rho) = \sum_{k=1}^{M} n_k \Bigg[ &\log\left(\binom{M-1}{k-1}\right) + \sum_{i=0}^{k-2} \log\left(\frac{(\psi - 1)\rho}{M-1} + i(C(\psi) - \rho)\right) \\ &+ \sum_{i=0}^{M-1-k} \log\left(\frac{(M - \psi)\rho}{M-1} + i(C(\psi) - \rho)\right) \\ &- \sum_{i=0}^{M-2} \log\left(\rho + i(C(\psi) - \rho)\right) \Bigg], \end{aligned}$$

and for $\rho \geq C(\psi)$ is equal to

$$l(\psi, \rho) = \sum_{k=1}^{M} n_k \Bigg[ \log \Bigg( (\rho - C(\psi))[1 - |k - \psi|]_+$$

$$+\,(1-\rho)\binom{M-1}{k-1}\left(\frac{\psi-1}{M-1}\right)^{k-1}\left(\frac{M-\psi}{M-1}\right)^{M-k}\Bigg)-\log(1-C(\psi))\Bigg].$$

The gradient for $\rho < C(\psi)$ is equal to

$$\begin{aligned}\frac{\partial l}{\partial\psi}(\psi,\rho) &= \sum_{k=1}^{M} n_k\Bigg[\sum_{i=0}^{k-2}\frac{\frac{\rho}{M-1}+iC'(\psi)}{\frac{(\psi-1)\rho}{M-1}+i(C(\psi)-\rho)}\\ &\quad+\sum_{i=0}^{M-1-k}\frac{-\frac{\rho}{M-1}+iC'(\psi)}{\frac{(M-\psi)\rho}{M-1}+i(C(\psi)-\rho)}-\sum_{i=0}^{M-2}\frac{iC'(\psi)}{\rho+i(C(\psi)-\rho)}\Bigg],\\ \frac{\partial l}{\partial\rho}(\psi,\rho) &= \sum_{k=1}^{M} n_k\Bigg[\sum_{i=0}^{k-2}\frac{\frac{\psi-1}{M-1}-i}{\frac{(\psi-1)\rho}{M-1}+i(C(\psi)-\rho)}\\ &\quad+\sum_{i=0}^{M-1-k}\frac{\frac{M-\psi}{M-1}-i}{\frac{(M-\psi)\rho}{M-1}+i(C(\psi)-\rho)}+\sum_{i=0}^{M-2}\frac{i-1}{\rho+i(C(\psi)-\rho)}\Bigg],\end{aligned}$$

and for $\rho \geq C(\psi)$ the gradient is equal to

$$\begin{aligned}&\frac{\partial l}{\partial\psi}(\psi,\rho)\\ &\quad=\sum_{k=1}^{M} n_k\Bigg[\frac{(\rho-C(\psi))(\mathbb{1}_{[k-1,k]}(\psi)-\mathbb{1}_{[k,k+1]}(\psi))-C'(\psi)[1-|k-\psi|]_+}{(\rho-C(\psi))[1-|k-\psi|]_+ + (1-\rho)\tbinom{M-1}{k-1}\left(\frac{\psi-1}{M-1}\right)^{k-1}\left(\frac{M-\psi}{M-1}\right)^{M-k}}\\ &\qquad+\frac{\frac{(k-1)(1-\rho)}{M-1}\tbinom{M-1}{k-1}\left(\frac{\psi-1}{M-1}\right)^{k-2}\left(\frac{M-\psi}{M-1}\right)^{M-k}}{(\rho-C(\psi))[1-|k-\psi|]_+ + (1-\rho)\tbinom{M-1}{k-1}\left(\frac{\psi-1}{M-1}\right)^{k-1}\left(\frac{M-\psi}{M-1}\right)^{M-k}}\\ &\qquad-\frac{\frac{(M-k)(1-\rho)}{M-1}\tbinom{M-1}{k-1}\left(\frac{\psi-1}{M-1}\right)^{k-1}\left(\frac{M-\psi}{M-1}\right)^{M-1-k}}{(\rho-C(\psi))[1-|k-\psi|]_+ + (1-\rho)\tbinom{M-1}{k-1}\left(\frac{\psi-1}{M-1}\right)^{k-1}\left(\frac{M-\psi}{M-1}\right)^{M-k}}\\ &\qquad+\frac{C'(\psi)}{1-C(\psi)}\Bigg],\\ &\frac{\partial l}{\partial\rho}(\psi,\rho)\\ &\quad=\sum_{k=1}^{M} n_k\frac{[1-|k-\psi|]_+ - \tbinom{M-1}{k-1}\left(\frac{\psi-1}{M-1}\right)^{k-1}\left(\frac{M-\psi}{M-1}\right)^{M-k}}{(\rho-C(\psi))[1-|k-\psi|]_+ + (1-\rho)\tbinom{M-1}{k-1}\left(\frac{\psi-1}{M-1}\right)^{k-1}\left(\frac{M-\psi}{M-1}\right)^{M-k}}.\end{aligned}$$

## Appendix C: G-test, bootstrap procedure

Denote by $(n_1, \ldots, n_M)$ numbers of observed responses, i.e., $n_k$ is the number of responses assigned to response category $k$ and $\sum_{k=1}^{M} n_k = n$. By $(p_1, \ldots, p_M)$ denote unknown probabilities of the response categories $1, \ldots, M$. We want to test

$$H_0 : \ (p_1, \ldots, p_M) \text{ are from the GSD}$$

against

$$H_1 : \ (p_1, \ldots, p_M) \text{ are not from the GSD.}$$

One should not use the chi-squared test in case of small numbers in selected cells, i.e., small $n_k$ for some $k \in \{1, \ldots, M\}$. We use a bootstrap version of the standard likelihood ratio test, i.e., the G-Test. The procedure is as follows:

1. Estimate probabilities of the response categories $(\hat{p}_1, \ldots, \hat{p}_M)$ using the maximum likelihood GSD estimator.
2. Calculate test statistic $T = \sum_{k=1}^{M} n_k \log(n_k/(n\hat{p}_k))$, where $0 \log(0/(n\hat{p}_k)) = 0$.
3. Generate $MC$ (for example, $MC = 10{,}000$) bootstrap samples of size $n$ from the distribution $(\hat{p}_1, \ldots, \hat{p}_M)$. Obtain $(m_1^r, \ldots, m_M^r)$, $r = 1, \ldots, MC$, where $m_k^r$ is the number of responses assigned to response category $k$ in the $r$-th bootstrap sample.
4. Estimate probabilities of the response categories $(\hat{q}_1^r, \ldots, \hat{q}_M^r)$ for every bootstrap sample $(m_1^r, \ldots, m_M^r)$ using the maximum likelihood GSD estimator.
5. Calculate bootstrap statistics

$$T_r = \sum_{k=1}^{M} m_k^r \log(m_k^r/(n\hat{q}_k^r)),$$

   where $0 \log(0/(n\hat{q}_k^r)) = 0$.
6. Calculate bootstrap $p$-value using the following equation

$$p = \frac{1}{MC} \sum_{r=1}^{MC} I(T_r \geq T),$$

   where $I(x)$ is one if $x$ is true or 0 if $x$ is false.

## Appendix D: Modified EPMF and GSD

To resolve the problem of empty cells for the empirical distribution, we can simply add 0.5 to all response category counts (cf. Pagano and Gauvreau (2018)), i.e.,

$$\forall k \in \{1, \ldots, M\}, \quad \hat{v}_k = \frac{n_k + 0.5}{n + \frac{M}{2}},$$

where $n_k$ is the response count of category $k$ in a bootstrap sample.

For the GSD it is enough to estimate parameters $\psi$, $\rho$ on the set $[1 + \epsilon_{\psi_d}(n), 5 - \epsilon_{\psi_u}(n)] \times [0 + \epsilon_{\rho_\mathrm{d}}(n), 1 - \epsilon_{\rho_\mathrm{u}}(n)]$, where $\epsilon_{\psi.}(n) > 0$, $\epsilon_{\rho.}(n) > 0$ and $\lim_{n \to \infty} \epsilon_{\psi.}(n) = \lim_{n \to \infty} \epsilon_{\rho.}(n) = 0$.

To define $\epsilon_{\psi.}(n)$ and $\epsilon_{\rho.}(n)$ we introduce a limit for the maximum probability any two response categories can add up to (and call it $p_{\max}$). Importantly, when assessing $p_{\max}$, we only take into account two most probable response categories. This can be formally written as follows:

$$p_{\max} = \max_{(i,j) \in \{1,\ldots,M\}^2 : i \neq j} P(U = i) + P(U = j) \tag{D1}$$

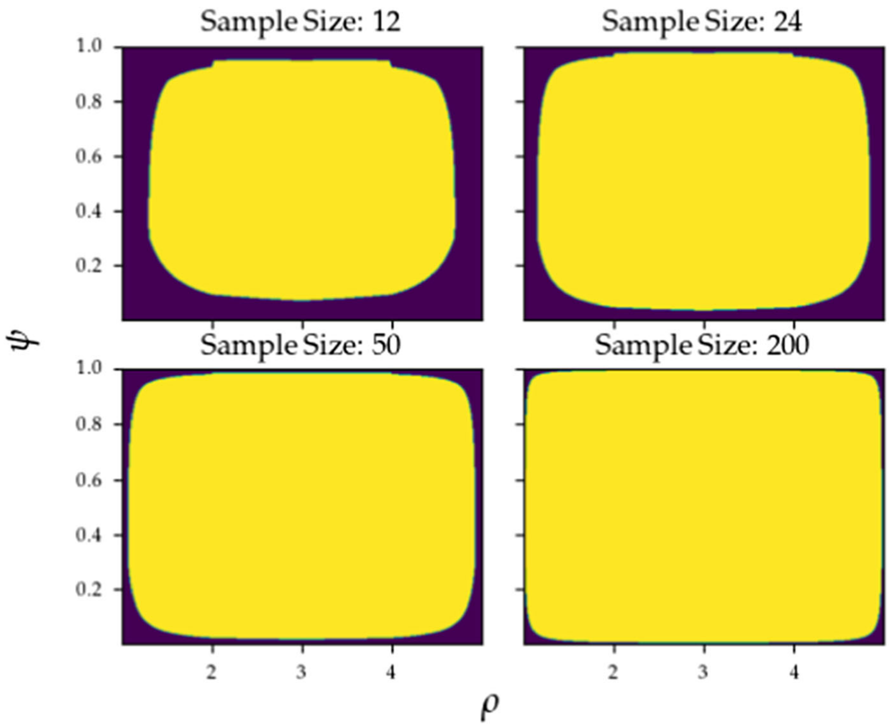


**Fig. 13** Boundary for $\psi$ and $\rho$ for a given sample size $n$ and $p_{\max} \leq 1 - \frac{1}{n}$. Yellow color marks $(\psi, \rho)$ pairs considered in the MLE algorithm

The final algorithm for fitting the GSD to a sample is as follows. Find such $(\hat{\psi}, \hat{\rho})$ that satisfies the following two criteria:

1. $p_{\max} \leq 1 - \frac{1}{n}$, where $p_{\max}$ is given by Eq. (D1) and $n$ is the sample size, and
2. the likelihood function has the maximum value.

An example of $\psi$ and $\rho$ ranges for different sample sizes and $M = 5$ is shown in Fig. 13.

## Appendix E: mulitidimensional case

Let us consider a multidimensional model of responses with $n$ raters (also referred to as subjects) and $m$ objects (also referred to as stimuli), i.e.,

$$U_{ij} = \psi_j + \epsilon_{ij}, \quad i \in 1, \ldots, n, \; j \in 1, \ldots, m$$

where $\epsilon_{ij} + \psi_j$ has the $\text{GSD}(\psi_j, \rho_i)$ distribution. In this model, every object (e.g., a video) has its own quality $\psi_j$ and every rater has their own confidence parameter $\rho_i$.

For numerical experiments, we generated 100,000 response matrices $U_{i,j}$ according to the following generative process: $\psi_j \sim \text{Uniform}(1, 5)$, $\rho_i \sim \text{Uniform}(0, 1)$. For each sample a probabilistic matrix factorisation $U_{ij} \; \sim GSD(\hat{\psi}_j, \hat{\rho}_i)$ done by MLE yields recovered estimates $\hat{\psi}_j$ and $\hat{\rho}_i$.

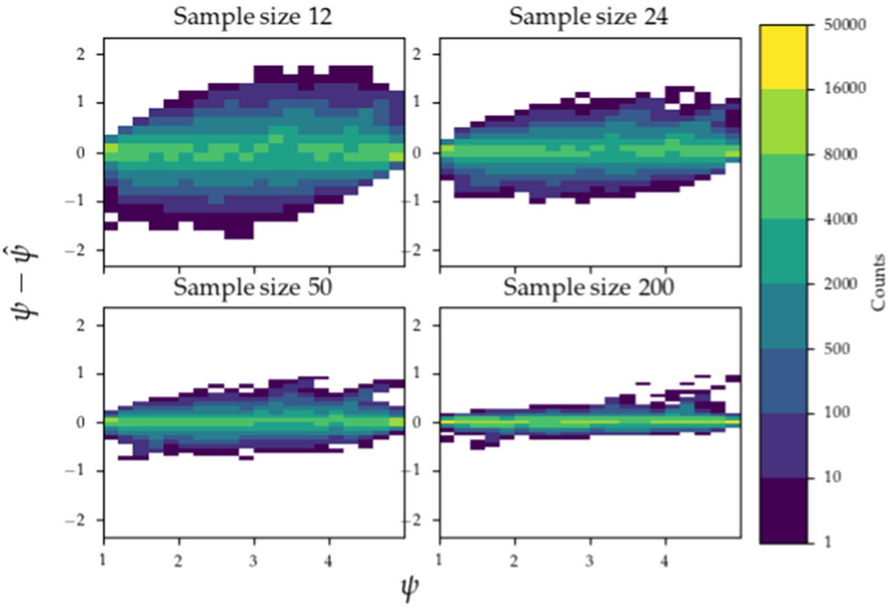


**Fig. 14** Estimation accuracy for $\psi$ estimation for the multidimensional model with $n$ subjects scoring $m = n$ objects

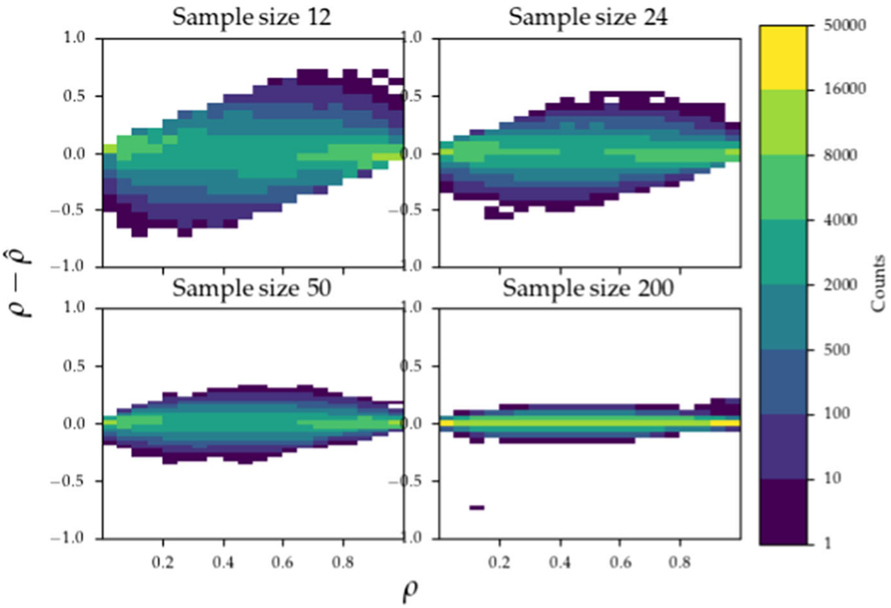


**Fig. 15** Estimation accuracy for $\rho$ estimation for the multidimensional model with $n$ subjects scoring $m = n$ objects

In Fig. 14 we present the estimation accuracy for a fixed $\psi$ of one object in a multidimensional numerical experiment for $n = m = 12, 24, 50, 200$. All other parameters where random (uniformly distributed). For estimation, we used the gradient based method, using the formulae for the gradient of GSD log-likelihood function from Appendix B. In Fig. 15 we present estimation accuracy for fixed $\rho$ of one rater in the same multidimensional numerical experiment. As one can see, our multidimensional estimator of GSD parameters is very accurate even for relatively small sample sizes. Notice that in case of $n = m = 200$, we estimate 400 parameters using the MLE, i.e., we estimate $\psi_1, \ldots, \psi_{200}, \rho_1, \ldots, \rho_{200}$. The accuracy of the estimator is higher if both $n$ and $m$ are larger.

## References

Agresti A (2002) Categorical data analysis, 2nd edn. Wiley, New York

Alwin DF, Baumgartner EM, Beattie BA (2018) Number of response categories and reliability in attitude measurement. J Surv Stat Methodol 6:212–239

Becker WE, Kennedy PE (1992) A graphical exposition of the ordered probit. Economet Theor 8:127–131. https://doi.org/10.1017/S0266466600010781

Chinen M (2021) Marginal effects of language and individual raters on speech quality models. IEEE Access 9:127,320-127,334. https://doi.org/10.1109/ACCESS.2021.3112165

Chinen M, Skoglund J, Hines A (2021) Speech quality estimation with deep lattice networks. J Acoust Soc Am 149(6):3851–3861. https://doi.org/10.1121/10.0005130

Coombes KR (2018) The beta-binomial distribution. https://cran.r-project.org/web/packages/TailRank/vignettes/betabinomial.pdf

Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman and Hall, New York

Gange SJ, Munoz A, Saez M et al (1996) Use of the beta-binomial distribution to model the effect of policy changes on appropriateness of hospital stays. J R Stat Soc Ser C (Appl Stat) 45:371–382

Griffiths DA (1973) Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. Biometrics 29:637–648

Hossfeld T, Heegaard PE, Varela M, et al (2018) Confidence interval estimators for MOS values. arXiv:1806.01126

Hoßfeld T, Seufert M, Naderi B (2021) On inter-rater reliability for crowdsourced qoe. In: 2021 13th International conference on quality of multimedia experience (QoMEX), pp 37–42. https://doi.org/10.1109/QoMEX51781.2021.9465382

ITU-R (2019) Recommendation 500-14: methodology for the subjective assessment of the quality of television images. ITU-R Rec. BT.500

ITU-T Study Group 12 (1998) ITU-T coded-speech database. http://handle.itu.int/11.1002/1000/4415

Janowski L, Pinson M (2015) The accuracy of subjects in a quality experiment: a theoretical subject model. IEEE Trans Multimedia 17(12):2210–2224. https://doi.org/10.1109/TMM.2015.2484963

Janowski L, Nawała J, Robitza W, et al (2019) Notation for subject answer analysis

Li Z, Bampis CG (2017) Recover subjective quality scores from noisy measurements. In: Data compression conference proceedings part F127767, pp 52–61. https://doi.org/10.1109/DCC.2017.26

Li Z, Bampis CG, Janowski L, et al (2020) A simple model for subject behavior in subjective experiments. In: Human vision and electronic imaging (HVEI) 2020. arXiv:2004.02067

Liddell TM, Kruschke JK (2018) Analyzing ordinal data with metric models: what could possibly go wrong? J Exp Soc Psychol 79:328–348. https://doi.org/10.1016/j.jesp.2018.08.009

Malott DL, Fulton BR, Rigamonti D et al (2017) Psychometric testing of a measure of patient experience in Saudi Arabia and the United Arab Emirates. J Surv Stat Methodol 5:398–408

McCullagh P, Nelder J (1989) Generalized linear models. Chapman and Hall, New York

Nawała J, Janowski L, Ćmiel B, et al (2020) Describing subjective experiment consistency by p-value p-p plot. In: Proceedings of the 28th ACM international conference on multimedia. ACM, New York, pp 852–861. https://doi.org/10.1145/3394171.3413749

Nawała J, Janowski L, Ćmiel B, et al (2022) Generalised score distribution: a two-parameter discrete distribution accurately describing responses from quality of experience subjective experiments. IEEE Trans Multimedia (Accepted)

Pagano M, Gauvreau K (2018) Principles of biostatistics. Chapman and Hall/CRC, New York

Perez P, Janowski L, Garcia N, et al (2021) Subjective assessment experiments that recruit few observers with repetitions (fowr). IEEE Trans Multimedia. https://doi.org/10.1109/TMM.2021.3098450

Pinson MH (2018) Its4s: a video quality dataset with four-second unrepeated scenes. NTIA Technical Memorandum 18-532. https://www.its.bldrdoc.gov/publications/details.aspx?pub=3194

Pinson MH, Janowski L (2014) Agh/ntia: a video quality subjective test with repeated sequences. Tech. Rep. NTIA Technical Memo TM-14-505, NTIA/ITS

Pinson M, Speranza F, Barkowski M, et al (2010) Report on the validation of video quality models for high definition video content. Video Quality Experts Group

Pinson MH, Janowski L, Pepion R et al (2012) The influence of subjects and environment on audiovisual subjective tests: an international study. IEEE J Select Topic Signal Process 6(6):640–651. https://doi.org/10.1109/JSTSP.2012.2215306

Prentice RL (1986) Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. J Am Stat Assoc 81:321–327